Audibility of time switching in dynamic binaural synthesis

Hoffmann, Pablo F.F.; Møller, Henrik

Published in:
Journal of the Audio Engineering Society

Publication date:
2005

Audio Engineering Society

Convention Paper 6326

Presented at the 118th Convention
2005 May 28–31 Barcelona, Spain

Audibility of Time Switching in Dynamic Binaural Synthesis

Pablo Faundez Hoffmann[1] and Henrik Møller[1]

[1] *Department of Acoustics, Aalborg University, Denmark*

Correspondence should be addressed to Pablo F. Hoffmann (pfh@acoustics.aau.dk)

ABSTRACT

In binaural synthesis, signals are convolved with head-related transfer functions HRTFs. In dynamic systems, the update is often done by cross-fading between signals that have been filtered in parallel with two HRTFs. An alternative to cross-fading that is attractive in terms of computing power is direct switching between HRTFs that are close enough in space to provide an adequate auralization of moving sound. However, direct switching between HRTFs does not only move the sound but may also generate artifacts such as audible clicks. HRTF switching involves switching of temporal characteristics (ITD) and spectral characteristics, and the audibility of these were studied separately. The first results, data on minimum audible time switch, MATS, are presented.

1. INTRODUCTION

Binaural synthesis is the part of binaural technology [1, 2] that aims at artificially generating three-dimensional sound by only using two audio channels. Head-related transfer functions HRTFs are the kernel of binaural synthesis. These transfer functions uniquely describe the direction-dependent sound transmission from a free field to the ears of a listener. HRTFs can be represented as a pair of linear and time-invariant systems or filters, where each filter is associated to one ear. In this sense, a stationary sound image can be simulated by convolving an HRTF with an anechoically recorded sound. The sound image is then perceived as coming from the direction related to the HRTF when reproduced through an adequately equalized playback chain e.g. through headphones. Dynamic binaural synthesis, which concerns synthesis of a sound field that changes over time e.g. as a result of moving sound sources, listener movements or head movements, represents a more complex scenario to implement. The increase in complexity has its origin in the fact that the convolution of the anechoic signal with the HRTFs becomes time-variant. HRTFs must be updated according to the spatial location of a sound source.

Since HRTFs refer to discrete directions, direct switching between HRTFs may result in a non-smooth movement, where the discrete steps can be perceived. Furthermore, the switching operation itself may result in artifacts, such as audible clicks. These problems are usually overcome by cross fading between signals that have been convolved with HRTFs from two or more directions. A disadvantage by this solution is that two or more convolutions have to be performed in parallel, thus, demanding extra computing power. Therefore it is worth considering, whether it would be possible to use HRTFs of such fine resolution that neither the discrete steps of the movements nor switching artifacts can be heard. This would require more HRTFs in the database, thereby exchanging demands to computing power with demands to memory. The present study is part of an investigation that aims at determining the spatial resolution needed for direct switching of HRTFs and assessing the feasibility of this solution in practical applications.

There are two important requirements related to direct switching of HRTFs that must be examined in detail. First, in order to make the sound transition to be perceived smooth and continuous, each switching step must be below the human audibility threshold. This threshold is given by the minimum audible angle (MAA) which was defined by Mills as the smallest detectable difference between the angles of two sound sources [3]. In his study, Mills reported MAA as low as 1° for sounds directly in front of the listeners. It was also shown that MAA increases to more than 10° as the sound was lateralized. MAAs of 4°-5° at 0° azimuth in the horizontal plane have been reported when using synthesized spatial sound presented through headphones [4].

The second aspect related to direct switching of HRTFs is the switching operation itself (also defined as commutation [5]). This means that even though the first aspect is satisfied, the differences between the HRTFs will cause a discontinuity in the signal at the moment of switching. These discontinuities, which are more generally described as artifacts, might become audible e.g. as 'clicks' and, therefore, they should also be below the human audibility threshold, here denoted a minimum audible switch (MAS). Note that when HRTFs are modeled as FIR filters only discontinuities occur. If IIR models are used then a problem concerning transients in the output signal must also be accounted for [6].

The differences between HRTFs can be separated into temporal characteristics i.e. interaural time differences (ITD), and magnitude spectra characteristics given by the minimum-phase transfer functions to each of the ears. Therefore, it can be assumed that, when direct switching between HRTFs is performed, the generation of artifacts strongly depends on how large the differences between their respective temporal and spectral properties are. This line of reasoning has been the background for investigating two audibility thresholds related to direct switching: a minimum audible time switch (MATS) and a minimum audible spectral switch (MASS). MATS is defined as the smallest pure time switching that causes an audible artifact. MASS accounts for the smallest pure spectral switching between HRTFs that causes an audible artifact.

The required resolution for direct HRTF switching is expected to be finer than the resolution that HRTFs are usually measured with [7, 8]. Interpolation algorithms can be implemented to construct the HRTF database with the necessary resolution [9, 10]. How fine the original measurements must be has been reported by [11].

After this discussion, it is worth noting that cross-fading and direct HRTF switching in a linearly interpolated database of HRTFs are mathematically the same. Cross-fading is expressed as

$$y(n) = x(n) * h_i(n) \cdot \alpha + x(n) * h_j(n) \cdot (1 - \alpha) \quad (1)$$

where $x(n)$ and $y(n)$ are the input (anechoic) and output signal respectively, α is the cross-fading factor that goes from 0 to 1, and $h(n)$ is the time representation of the HRTFs - also denoted head-related impulse responses HRIR - where the sub-indexes i and j denote different directions. Since $x(n)$ is common to both convolutions, eq. 1 can be rearranged to

$$y(n) = x(n) * [h_i(n) \cdot \alpha + h_j(n) \cdot (1 - \alpha)] \quad (2)$$

which is the mathematical expression of a single convolution with an interpolated HRTF (linearly interpolated in the time domain). This means that if the HRTF interpolation is applied in real-time there is a potential for saving computing power even without extra memory requirements. Also for this solution, there is a need to know how large are the steps that can be allowed between HRTF updates.

The present work will report an experiment performed in order to obtain MATS values for several spatial locations. Their implications on dynamic binaural synthesis are posteriorly discussed.

2. EXPERIMENTAL METHOD

The main purpose of this study was to estimate MATSs by setting up a system where direct switching between the temporal characteristics of HRTFs was applied. A psycho-acoustical experiment was performed where sound signal, direction, and switch rate were varied. Spectral characteristics of the HRTFs remained unchanged during the switching operation.

2.1. Stimuli

Two different types of sound signals were used for this experiment: A 20 Hz - 9 kHz band-pass-filtered pink noise and a 1 kHz tone. The pink noise was chosen to represent general broadband signals, while the pure tone was chosen as a signal that would probably give less masking of the clicks. To render directional sound, the pink noise was filtered with a set of HRTFs selected from thirteen HRTF sets corresponding to directions summarized on Table 1. Directions are given as (azimuth, elevation) in a polar coordinate system with horizontal axis and left-right poles. -90° is to the left, 0° elevation in front and 90° elevation above. Five directions were selected in the median plane with a resolution of about 45°. Three directions were chosen on a cone of confusion to the left ((58°, 0°), (46°, 90°) and (54°, 180°)), and three on a similar cone to the right ((-56°, 0°), (-46°, 90°) and (-54°, 180°)). Directions were chosen to have the same ITD rather than being on a geometrical cone, thus their azimuth varies with elevation. A small asymmetry of the head is reflected in a small difference between sides in the azimuth at 0° elevation. Directions directly to the sides, corresponding to (90°, 0°) and (-90°, 0°) were also selected. The 1 kHz tone was filtered with a subset of this HRTF set corresponding to (0°, 0°), (90°, 0°), (0°, 180°), and (-90°, 0°). This gave a total of 17 different binaural stimuli.

HRTFs were obtained from a database of measurements made using an artificial head with a high directional resolution [7, 12]. The directions were chosen to cover an acceptable range of the upper half of the sphere and also to cover a wide ITD range. ITD values were derived from the interaural group delay difference of the excess-phase components of the HRTFs evaluated at 0 Hz [13]. HRTFs were implemented as minimum-phase FIR filters with a length of 1.5 ms (72 taps at 48 kHz sampling frequency). This decision was made considering that minimum-phase representation of HRTFs plus ITD does not perceptually differ from the original as long as the ITD is determined correctly [14]. It has also been demonstrated in a previous experiment that a length of 1.5 ms is sufficient to avoid audible effects of the truncation [15]. The DC value of each HRTF was calibrated to 0 dB gain [1]. Implementation of the ITD was done by inserting it as a pure integer delay.

Stimuli were played back by a computer through a D/A converter of 16 bit resolution at a sampling frequency of 48 kHz. In order to simulate a continuous sound, stimuli of 5 s for the pink noise and 1 s for the tone were looped and care was taken to avoid audible artifacts at the moment of looping. The output signal was fed to a pair of Beyerdynamic DT-990 headphones. Two minimum-phase filters were applied to the stimuli in order to compensate for the left and right headphone transfer functions respectively. The design of the equalization filters was based on headphone transfer functions (PTF) measured in a block ear canal obtained on 25 subjects. Five PTFs were obtained from each ear on each subject. PTFs were averaged on sound power basis and a minimum-phase representation of the inverse of the averaged PTF was computed for each ear. Details of the measurement and equalization technique are given in [16]. Fade-in and -out ramps of 10 ms were also applied. All stimuli were produced off-line and

Direction (azimuth, elevation)	ITD (μs)	Approximated sample index
(0°, 0°)	2.3	0
(0°, 44°)	0.7	0
(0°, 90°)	0.7	0
(0°, 136°)	0.8	0
(0°, 180°)	2.8	0
(58°, 0°)	-439.7	21
(46°, 90°)	-433.4	21
(54°, 180°)	-430.9	21
(-56°, 0°)	430.8	21
(-46°, 90°)	435.3	21
(-54°, 180°)	430.6	21
(90°, 0°)	-621.5	30
(-90°, 0°)	624.6	30

Table 1: Directions and ITD values of the HRTFs used in the listening experiment. Azimuth and elevation are given in polar coordinates where the poles are assigned to left and right. The approximated sample index corresponds to the number of zero-valued samples inserted at the beginning of the contralateral impulse response of the HRTFs to simulate the ITD.

stored on a hard disk. The gain of the system was calibrated so as to simulate a free-field sound pressure level of 72 dB.

2.2. Time Switching Implementation

To implement time switching, a variable digital delay was applied in real-time to the binaural stimuli in a back-and-forth modality. In this way, a single stimulus presentation can be described as a continuous $ABABAB$ ... sequence. State A corresponded to a non-delayed part of the stimulus and state B corresponded to the consecutive part of the stimulus but delayed. Time switching was operated at two different rates - 100 Hz and 50 Hz. The delay was applied diotically (which simulates an alternate change in the propagation delay). The reason for not introducing the delay as an ITD was that it had been observed in pilot experiments that this makes it difficult to discriminate between switching artifacts and directional movements.

Pilot experiments also showed that MATSs were generally below one sample at a sampling frequency of 48 kHz. Therefore, time switching was implemented by using FIR fractional delay filters [17]. Filter coefficients were calculated by using the Lagrange interpolation design technique and the order of the filters was set to 11. This filter order ensured that the filters had a flat frequency response and constant phase delay in the frequency range of the stimuli. Since these filters needed to be exchanged during operation, a table lookup method was used where the coefficients of all the required fractional delay filters were calculated and stored in advance [18]. In this sense, during the time switching operation for a switch modality AB, the appropriate fractional delay filter was retrieved from memory and convolved with the signal in real-time. The inherent delay of the fractional delay filters of half their length was compensated for by applying this as a constant delay to state A. Integer parts of the delays were implemented separately by simply jumping the actual number of samples backward or forward in the signal from the current sample at the moment of switching.

2.3. Subjects

21 paid subjects participated in this listening experiment. Their ages ranged from 20 to 31. The panel of subjects consisted of 10 males and 11 females. Normal hearing was checked by means of an audiometry (hearing levels $\leq$ 10 dB, at octave frequencies between 250 and 4 kHz, and a hearing level $\leq$ 15 dB at 8 kHz).

2.4. Psychometric Method

The listening experiment was conducted by using the method of adjustment. The advantage of this method is that it requires more active participation of the subject, thus, increasing his/her concentration and reducing boredom. This method is also relatively fast to carry out.

When stimuli were presented, subjects sat in a quiet room in front of a screen. The screen displayed a graphical interface with a slider and a push button (labeled OK). Subjects could control the slider with a mouse. Fig. 1 shows the graphical interface presented to the subjects. The slider controlled the amount of delay applied to the signal. As the slider was moved up and down the delay was increased and decreased respectively. For estimating MATS, subjects were asked to find the lowest position of the slider where they could still hear the clicks on the signal. This required the subjects to move the slider up and down several times. Subjects were encouraged to perform the task as fast as they could. When the threshold was determined the push button was pressed. After a silence interval of 2 s a new stimulus was presented.

The minimum time switching used by the set-up was a tenth of a sample, which corresponds to about 2.1 μs. As the subject moved the slider upwards, the time switching was incremented logarithmically with 20 steps per decade, each step thus being 12.2%. The maximum time switching was set to 4.2 ms. In number of samples, delay values ranged from 0.1 to 199.53 samples giving a scale of 67 different delays.

The scale of delays was contained within a frame equal to half the length of the slider bar. The position of the frame along the slider bar was randomized. All delay values below the lower end of the frame were set to 0, and all values above the upper end of the frame were set to equal the maximum delay of the scale. This was done to ensure that the threshold position varied along the slider bar. Furthermore, a potential bias from response criterion based on visual cues e.g. distance of the slider from the bottom, was believed to be minimized. The initial position of the slider was also randomly selected at either the bottom or the top of the slider bar. This ensured that the slider position was at a clear distance from threshold at the beginning of each trial.

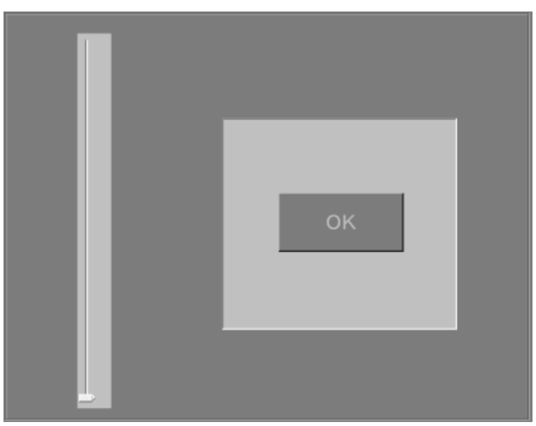

Fig. 1: Graphical interface presented to subjects during the experiment.

2.5. Experimental Design

The listening experiment consisted of a three-stage procedure:

- Hearing test and familiarization session,
- Practice session,
- Main experiment.

Each stage was carried out on different days, the main experiment, however, on three days. For the familiarization session subjects were initially provided with written instructions. After that, subjects were provided with an oral explanation just in case that some aspects of the written instructions were not fully understood. Then, subjects were presented with a block of trials. Here, as well as in the practice session and the main experiment, one block consisted of all sound stimuli (17 trials). The order in which the stimuli were presented was random, and switch rates (either eight times 50 Hz and nine times 100 Hz or vice versa) were randomly assigned to the stimuli. The practice session consisted of 3 blocks. Responses from practice sessions were not used for analysis. In the main experiment, each combination of stimulus and switch rate was repeated five times, which gives ten blocks with a total of 170 responses for each subject. The blocks were given on three days with three, three, and four blocks on a day. After each block a break of 5 - 7 minutes was given to the subjects.

3. RESULTS

No responses were above the highest time switch allowed by the system, and 0.2% (8) of the total number of responses fell below the smallest time switch allowed. These eight responses were considered invalid and were excluded from the analysis. Since data appeared to better represent normal

		Stimulus			
		Pink Noise		Tone	
	Sound	Switch Rate		Switch Rate	
ITD	Direction	100 Hz	50 Hz	100 Hz	50 Hz
0 μs	(0°, 0°)	5.3	6.1	5.7	6.8
	(0°, 44°)	5.0	5.6		
	(0°, 90°)	5.3	6.1		
	(0°, 136°)	5.3	6.1		
	(0°, 180°)	7.1	8.4	4.2	5.3
-437.5 μs	(58°, 0°)	7.3	9.1		
	(46°, 90°)	7.0	8.1		
	(54°, 180°)	6.4	7.4		
437.5 μs	(-56°, 0°)	5.7	6.5		
	(-46°, 90°)	6.3	7.0		
	(-54°, 180°)	5.7	6.2		
-625 μs	(90°, 0°)	8.5	9.4	3.6	4.5
625 μs	(-90°, 0°)	6.2	6.9	3.7	4.7

Table 2: Mean MATS values calculated across subjects. Thresholds are expressed in time units (μs).

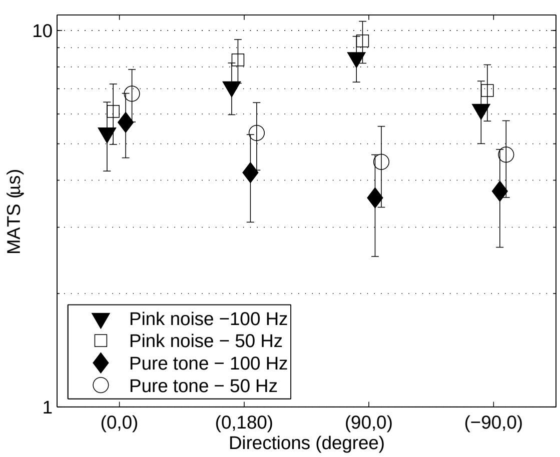

Fig. 2: MATS averaged across subjects for different sound stimulus and switch rate as function of azimuth changes. Error bars indicate standard error of the mean (s.e.m.).

distributions on a logarithmic scale than on a linear scale, all statistics was carried out on log(MATS). Individual MATSs were calculated as the mean of five repetitions. Means, standard deviations and standard errors of the means were then calculated for the group. For the presentation in figures and tables, data were transformed back to the linear scale. Results are summarized in Table 2. Fig. 2 shows MATSs for the directions where data exist for both pink noise and tone. Fig. 3 shows MATSs for pink noise arranged by locations at the different cones of confusion.

MATSs obtained for a switch rate of 100 Hz were all lower than MATSs obtained for a switch rate of 50 Hz, although the differences are not large and hardly statistically significant for each individual direction and signal. At the side locations and at the rear, lower MATSs were found for the pure tone than for the pink noise. The lowest MATSs obtained were 3.6 μs for the pure tone located at (90°, 0°), and 5.0 μs for pink noise located at (0°, 44°). Higher MATSs were obtained at side locations for pink noise and straight ahead of the listener for the pure tone. The highest MATSs for pink noise and pure tone were 9.4 μs at (90°, 0°) and 6.8 μs at (0°, 0°).

4. DISCUSSION

When an artifact such as a discontinuity is introduced to a time signal, this constitutes an additional signal which is heard as a click. The reason is because all the energy of this additional signal, which is spread all over the frequency range, is concentrated in a very short time interval. A higher switch rate will increase the number of artifacts per time unit in the signal, and therefore, will increase the probability for them to be audible. That would explain why MATS are lower at a higher switch rate.

A pure tone is a very narrow-band signal and it is assumed to offer less masking of the artifacts created at the moment of switching. Therefore, the audibility thresholds are also assumed to be lower than those for a signal with a more broadband frequency content. This is supported by our findings as can be observed from Fig. 2. MATSs for the pure tone are in general lower than those for pink noise. The only difference was observed at the forward location where, for both switch rates, MATSs for the two signals are similar. This situation is being analyzed since it is not fully understood.

MATSs obtained for pink noise tend to increase as the sound location moves from straight ahead to the left side of the listener, as can be observed when comparing Fig. 3(a) with Fig. 3(b) and Fig. 3(c). Why this is not happening with the same order of magnitude to the right side remains to be explained. Observations on the directions corresponding to the cones of confusion with an ITD of ± 437.5 μs show that MATSs seem to be independent of changes in the elevation of the sound source. This is not the case for directions in the median plane where MATSs from the rear differ from the others.

The MAA directly in front of the listener is around 1°. An angular step of 1° has an associated ITD change of around 7.5 μs. Therefore, MATSs seem to put higher demands to the spatial resolution, since it was shown that a time switching between 5.3 - 5.7 μs is enough to be audible. When sound is coming from the sides, a step of 6.2 μs would be allowed if we consider the lowest MATS obtained for pink noise at (90°, 0°). This time corresponds to an angular step of about 2°-3°.

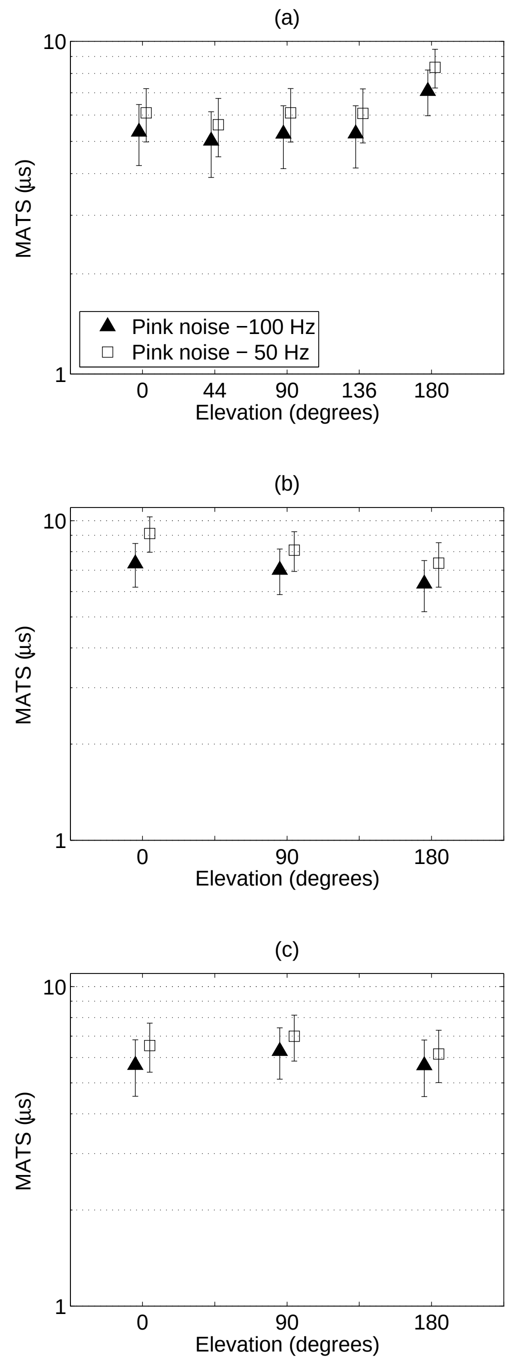

Fig. 3: MATS values across subjects for three cones of confusion, (a) 0 μs; (b)-437.5 μs; (c) 437.5 μs. Error bars indicate standard error of the mean (s.e.m.).

Since MAAs around these locations are greater than 10°, it can be seen that at the sides MATSs also determine the directional resolution.

MAAs have usually been obtained by using real sources. This means that the detection of changes in the sound location is based on temporal and spectral cues. Therefore a direct comparison between MATS and MAA might not be the most adequate but can only be used to get an idea of the requirements that each threshold imposes for the implementation of a dynamic system. Future experiments will be conducted in order to estimate MASS and also to assess audibility of angular differences attributed to temporal and spectral cues separately for several directions and for azimuth as well as elevation changes.

5. ACKNOWLEDGMENTS

The authors wish to thanks economic support from the Danish Technical Research Council.

6. REFERENCES

[1] D. Hammershøi and H. Møller. *Communication Acoustics*, chapter in Jens Blauert (edt.) : Binaural Technique, Basic Methods for Recording, Synthesis and Reproduction, pages 223–254. 2005. Springer Verlag. InPress.

[2] H. Møller. Fundamentals of binaural technology. *Applied Acoustics*, 36:171–128, 1992.

[3] A. W. Mills. On the minimum audible angle. *J. Acoust. Soc. Am.*, 30(4):237–246, December 1958.

[4] R. L. McKinley and M. A. Ericson. *Binaural and Spatial Hearing in Real and Virtual Environments*, chapter in R. H. Gilkey, T. R. Anderson (edt.) : Flight Demonstration of a 3-D Auditory Display, pages 683–699. Lawrence Erlbaum Associates, 1997. ISBN: 0-8058-1654-2.

[5] Jean-Marc Jot, V. Larcher, and O. Warusfel. Digital signal processing in the context of binaural and transaural stereophony. In *98th AES Convention*, Paris, France, February 25-28 1995. Preprint 3980.

[6] V. Välimäki and T. I. Laakso. Suppression of transients in variable recursive digital filters with a novel and efficient cancellation method. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 46(12), Dec. 1998.

[7] P. Minnaar, F. Christensen, S. K. Olesen, B. P. Bovbjerg, and H. Møller. Head-related transfer functions measured with a high directional resolution. *J. Audio Eng. Soc.* Submitted.

[8] W. G. Gardner and K. D. Martin. HRTF measurements of a kemar. *J. Acoust. Soc. Am.*, 97(6):3907–3908, 1995.

[9] F. Christensen, H. Møller, P. Minnaar, J. Plogsties, and S. K. Olesen. Interpolating between head-related transfer functions measured with low directional resolution. In *107th AES Convention*, New York, USA, September 24-27 1999. Preprint 5047.

[10] K. Hartung and J. Braasch. Comparison of different methods for the interpolation of head-related transfer functions. In *Proceedings of the 16th Audio Engineering Society (AES International Conference on Spatial Sound Reproduction*, Rovaniemi, Finland, 1999.

[11] P. Minnar, J. Plosgties, and F. Christensen. The directional resolution of head-related transfer functions required in binaural synthesis. *J. Audio Eng. Soc.* Submitted.

[12] B. P. Bovbjerg, F. Christensen, P. Minnaar, and X. Chen. Measuring the head-related transfer functions of an artificial head with a high directional resolution. In *109th AES Convention*, Los Angeles, California, USA, September 22-25 2000. Preprint 5264.

[13] P. Minnaar, J. Plogsties, S. K. Olesen, F. Christensen, and H. Møller. The interaural time difference in binaural synthesis. In *108th AES Convention*, Paris, France, February 19-22 2000. Preprint 5133.

[14] P. Minnaar, H. Møller, J. Plogsties, S. K. Olesen, and F. Christensen. The audibility of all-pass components in binaural synthesis. *J. Audio Eng. Soc.* In Preparation.

[15] J. Sandvad and D. Hammershøi. What is the most efficient way of representing HTF filters? In *Proceedings of Nordic Signal Processing Symposium*, NORSIG '94, Lesund, Norway, June 2-4 1994.

[16] H. Møller, D. Hammershøi, C. B. Jensen, and M. F. Sørensen. Transfer characteristics of headphones measured on human ears. *J. Audio Eng. Soc.*, 43(4), April 1995.

[17] T. I. Laakso, V. Välimäki, M. Karjalainen, and U. K. Laine. Splitting the unit delay. *IEEE Signal Processing Magazine*, 1996.

[18] V. Välimäki and T. I. Laakso. *Nonuniform Sampling Theory and Practice*, chapter in Farokh Marvasti (edt.) : Fractional Delay Filters - Design and Applications, pages 835–895. Kluwer Academic/Plenum Publishers, 2001.